ASSIGNMENT 3:

PREDICTING FUTURE OUTCOMES

TUNA TUNCAY

LSE Data Analytics Career Accelerator
Spring'22

TABLE OF CONTENTS

INTRODUCTION

PROBLEM DEFINITION:

Turtle Games would like to improve their sales performance. Based on their historic sales data and customer reviews received, below report is prepared to highlight key facts and come up with areas to focus on more.

METHADOLOGY

DATA CLEANING & THOUGHT PROCESS:

1. We are provided 2 data set. One includes the detail of remuneration, spending score, product ID and loyalty points. All data has been provided in CSV files. Second includes the sales amount for each product in different platform for 3 regions: Europe (EU), North America (NA) and Global.
2. Two files can be merged on product ID if needed.
3. Spending score and loyalty points from first file are used to cluster customer segments.
4. Second file is used to analyse sales volume and sales prediction.
5. Data cleaning process handled in Python and R. There were no missing values in both CSV files. I was concerned and worked on duplicated values and outliers.
6. During the analysis, I followed a progressive analysis style, meaning maintained multiple Jupiter notebooks and R-Studio files.

KEY FINDINGS

ACCUMULATION OF LOYALTY POINTS:

To understand how loyalty points are accumulated, I build several prediction models. For simplicity, initial models created as simple linear regression models (lrm) among 4 quantitative variables I have.

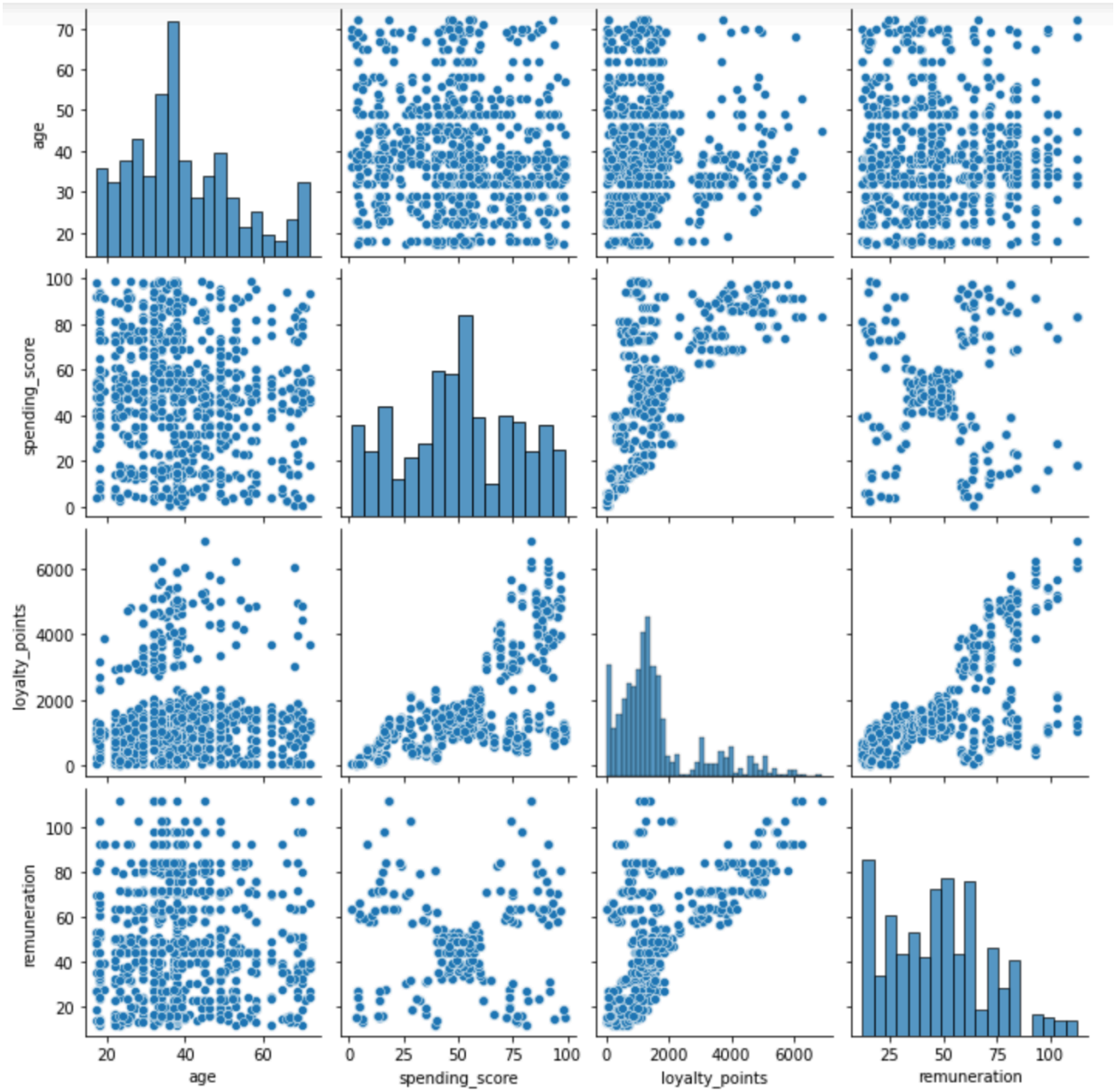
- Loyalty points ~ spending score
- Loyalty points ~ remuneration
- Loyalty points ~ age

Initial check with the pair plot suggested there is a positive relationship between first 2 but not clear relationship between loyalty & age. Model finding also suggested similar. While R-squared values are 45.2% and 37.9% for the first 2 respectively, it is only 2% for the last one.

I rebuilt the model with clean and no outlier data set to check whether there is an improvement, however the results for R-squared were even lower for all. Therefore, I decided to move forward with the dataset including outliers.

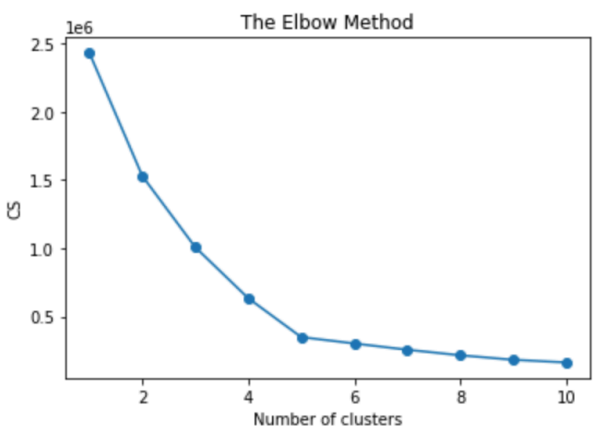
Next, multiple lrm is built. Keeping loyalty points as my dependent variable, I introduced the other variables (age, remuneration and spending) as my independent variables. Model R-squared increased to 83.9% with this method. Since I know from previous analysis as age and loyalty point are not correlated, I check exclusion of age from my model. R-squared slightly dropped by ~1% to %82.7.

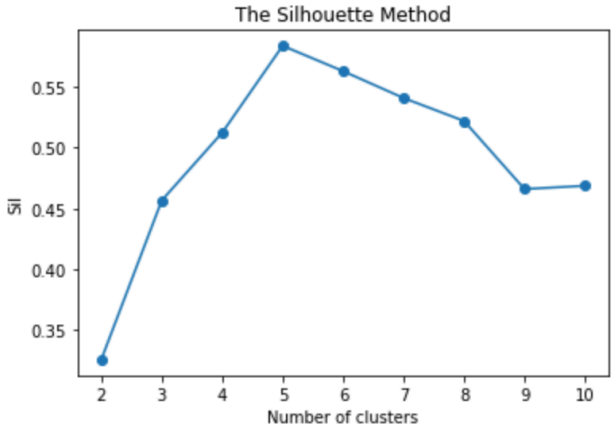
Loyalty point prediction can be made by the multiple lrm using remuneration and spending score.

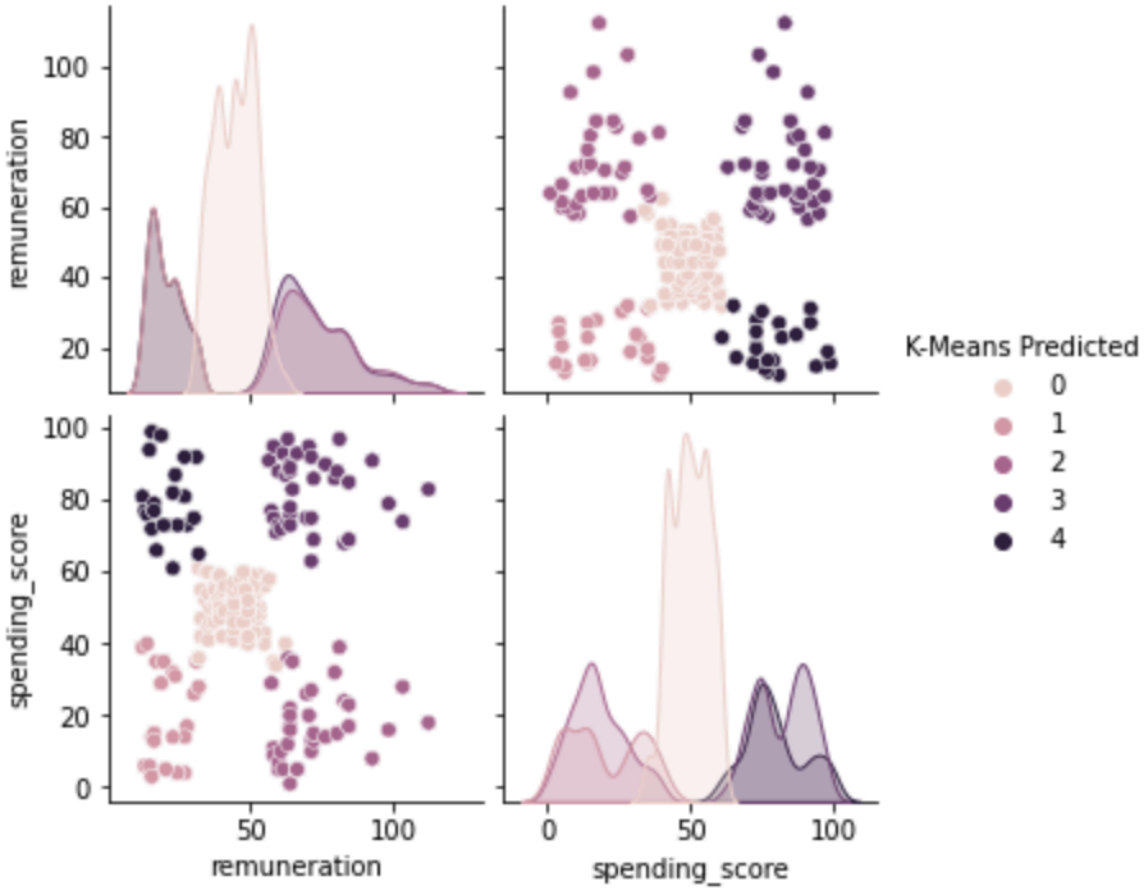


CUSTOMER SEGMENTATION:

I worked on customer segmentation based on remuneration and spending score. Using elbow and silhouette methods, created k-clusters. Due to the nature of the data, decided to move forward with 5 clusters which was also suggestion from earlier mentioned methods.


The Elbow Method
1e6
CS
Number of clusters


The Silhouette Method
Sil
Number of clusters


remuneration
spending_score
K-Means Predicted
0
1
2
3
4

Below is the size of the clusters in my dataset:

	remuneration		spending_score	
	count	mean	count	mean
K-Means Predicted				
AveSpending_and_AveRemuneration	774	44.418786	774	49.529716
HighSpending_and_HighRemuneration	356	73.240281	356	82.008427
HighSpending_and_LowRemuneration	269	20.353680	269	79.416357
LowSpending_and_HighRemuneration	330	74.831212	330	17.424242
LowSpending_and_LowRemuneration	271	20.424354	271	19.763838

Next, I merged my cluster data frame with original data set on renumeration and spending score to bring product ids. Due to duplicated data (same product can be sold in different platform and recorded in data frame as same remuneration, spending score and product id) merge function enlarged my dataset. I get rid of this problem by deleting repeated rows in my merged data frame. This changed my clusters a bit. I lost some data but still the main group proportions remained same.

	Remuneration		Spending_Score	
	count	mean	count	mean
Customer_Segments				
AveSpending_and_AveRemuneration	690	44.352493	690	49.547826
HighSpending_and_HighRemuneration	318	73.642704	318	81.996855
HighSpending_and_LowRemuneration	241	20.374108	241	79.269710
LowSpending_and_HighRemuneration	295	75.128678	295	17.461017
LowSpending_and_LowRemuneration	243	20.439259	243	19.860082

I used merged data frame to identify which products are purchased most for each cluster. This can help me target those segments with identified products. Below are the most purchased products for average earner & average spenders:

```
# HighSpending_and_HighRemuneration
df2_clusters[df2_clusters['Customer_Segments'] == 'HighSpending_and_HighRemuneration'].Product.value_counts().head(2
```

```
8923     5
10241    5
4399     5
10232    5
9080     5
4477     5
2518     5
2795     5
8962     5
6504     5
4415     5
3403     5
2457     5
10281    5
4619     4
4692     4
7573     4
11056    4
```

SENTIMENT ANALYSIS:

Data has reviews and summaries for sentiment analysis. I cleaned my data using word tokenization and elimination of stop words. Originally, there were 39 duplicated reviews and summaries but when I check them with product IDs realized that they are all for different products. Only 1 duplicated value (same row for product, review, and summary) deleted.

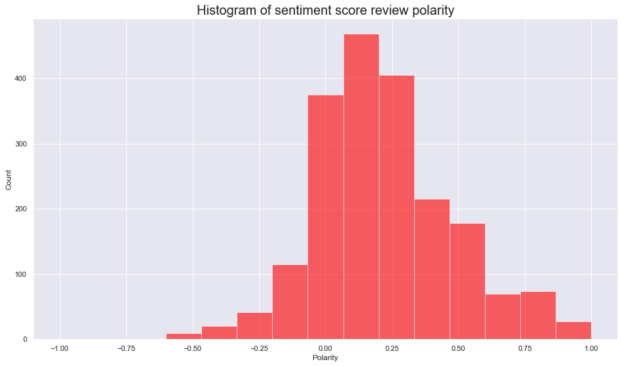
Word clouds can be found below. People are mentioning about products mostly with positive words.

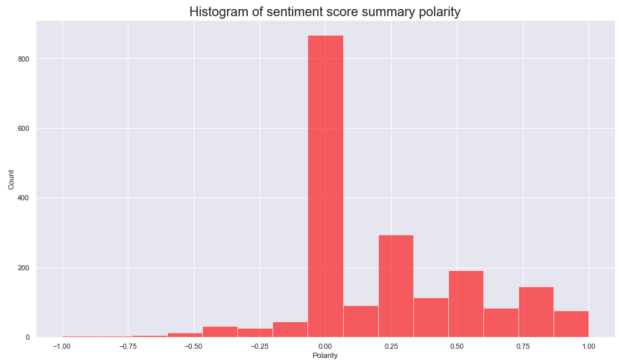
Word Cloud for Reviews

Word Cloud for Summary

Next, polarity analysis is conducted on reviews and summary. Polarity scores are positive for both.

	review_polarity	summary_polarity
count	1999.000000	1999.000000
mean	0.217697	0.219426
std	0.262567	0.335675
min	-1.000000	-1.000000
25%	0.049777	0.000000
50%	0.182179	0.062500
75%	0.361056	0.452500
max	1.000000	1.000000





Later, I try to come up with meaningful clusters based on polarity. My expectation was grouping them as perfect positive sentiment, perfect negative sentiment, neutral etc. Clusters suggested does not seem to be meaningful for me. Therefore, I stopped clustering analysis here.

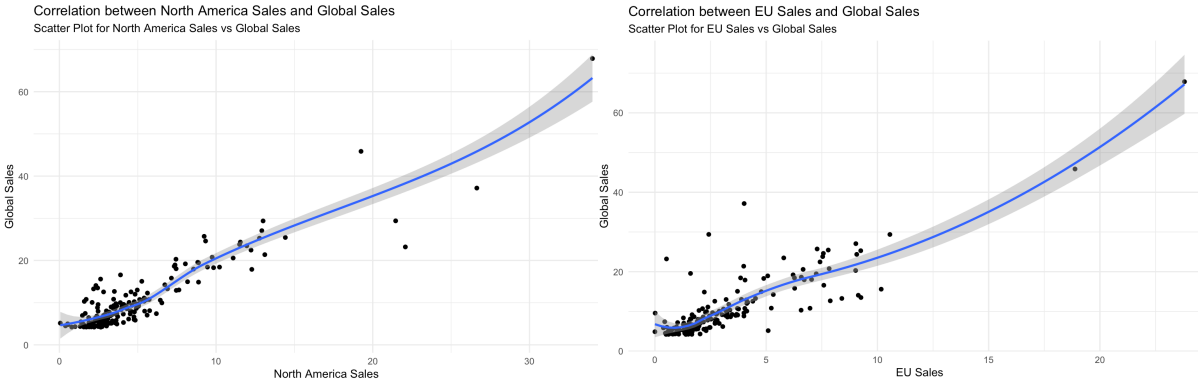
Lastly, I merged my data frame from week 2 to check sentiment polarity for different customer segments. They are not much different than each other.

Customer_Segments	Remuneration	Spending_Score	Review_Polarity	Summary_Polarity
AveSpending_and_AveRemuneration	44.352493	49.547826	0.203867	0.225120
HighSpending_and_HighRemuneration	73.642704	81.996855	0.233082	0.219954
HighSpending_and_LowRemuneration	20.374108	79.269710	0.219407	0.209188
LowSpending_and_HighRemuneration	75.128678	17.461017	0.235406	0.221597
LowSpending_and_LowRemuneration	20.439259	19.860082	0.215330	0.210670

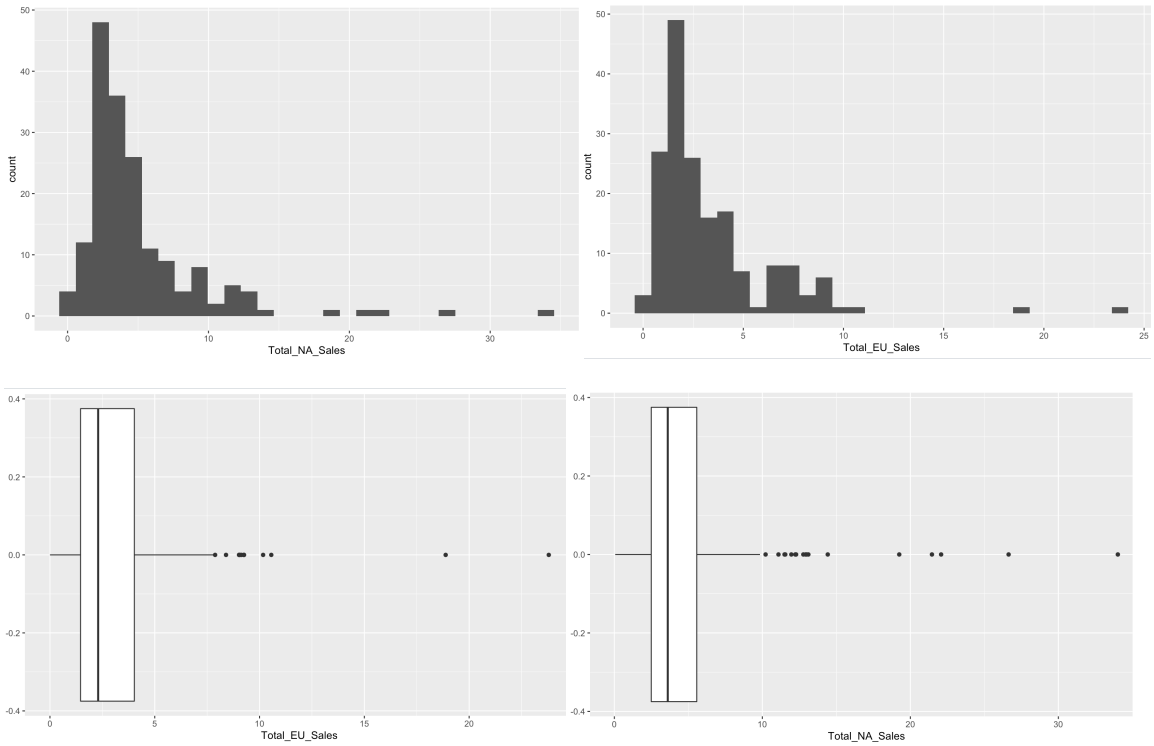
SALES ANALYSIS:

Sales analysis has been made with a focus on product detail. I consolidated my data with a group by function on product to eliminate the duplicate rows indicating sales from different platforms for each product. Then, I visualise my data on several plots for further analysis.

Scatter plot is used to understand the correlation between sales in different regions and their contribution to global sales. Both have positive correlation with global sales.



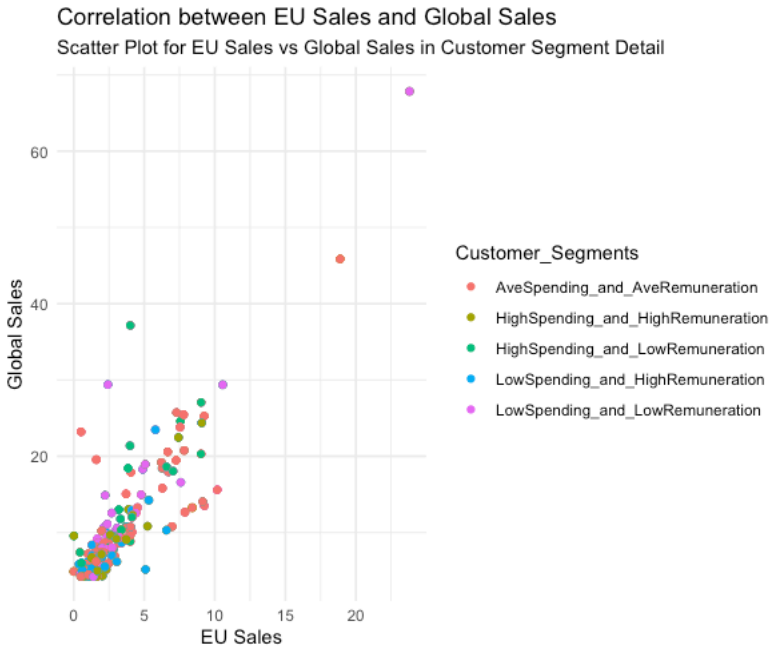
Box plots and histograms showed me data is right skewed and have outliers.



By using the data from histogram and box plot, I identified each regions outlier sales products. Below you can find product id's that are outlier in all sales regions.

	Product	count	Total_EU_Sales	Total_NA_Sales	Total_Global_Sales
	<fct>	<int>	<dbl>	<dbl>	<dbl>
1	107	1	23.8	34.0	67.8
2	195	1	10.6	13	29.4
3	231	1	9.03	12.9	27.1
4	515	5	18.9	19.2	45.9
5	876	4	9.25	12.8	25.3
6	979	4	9.07	11.5	24.4

Lastly, I merged customer segments with sales data using product ids. That helped me create scatter plots with customer segment breakdown. Realizing that highest sales from EU and NA comes from low spending and low renumeration group surprised me.



Histograms, Q-Q plots, Shapiro Tests and skewness and kurtosis of the sales data suggest that data is not normally distributed. This is prerequisite for correlation check. However, assumption of normality could be ignored given a large enough data set because Central Limit Theorem.

Below are the results for the correlation between sales data. Global sales are highly correlated with NA and EU sales.

```
                  Total_EU_Sales Total_NA_Sales Total_Global_Sales
Total_EU_Sales          1.0000000      0.6209317          0.8486148
Total_NA_Sales          0.6209317      1.0000000          0.9162292
Total_Global_Sales      0.8486148      0.9162292          1.0000000
```

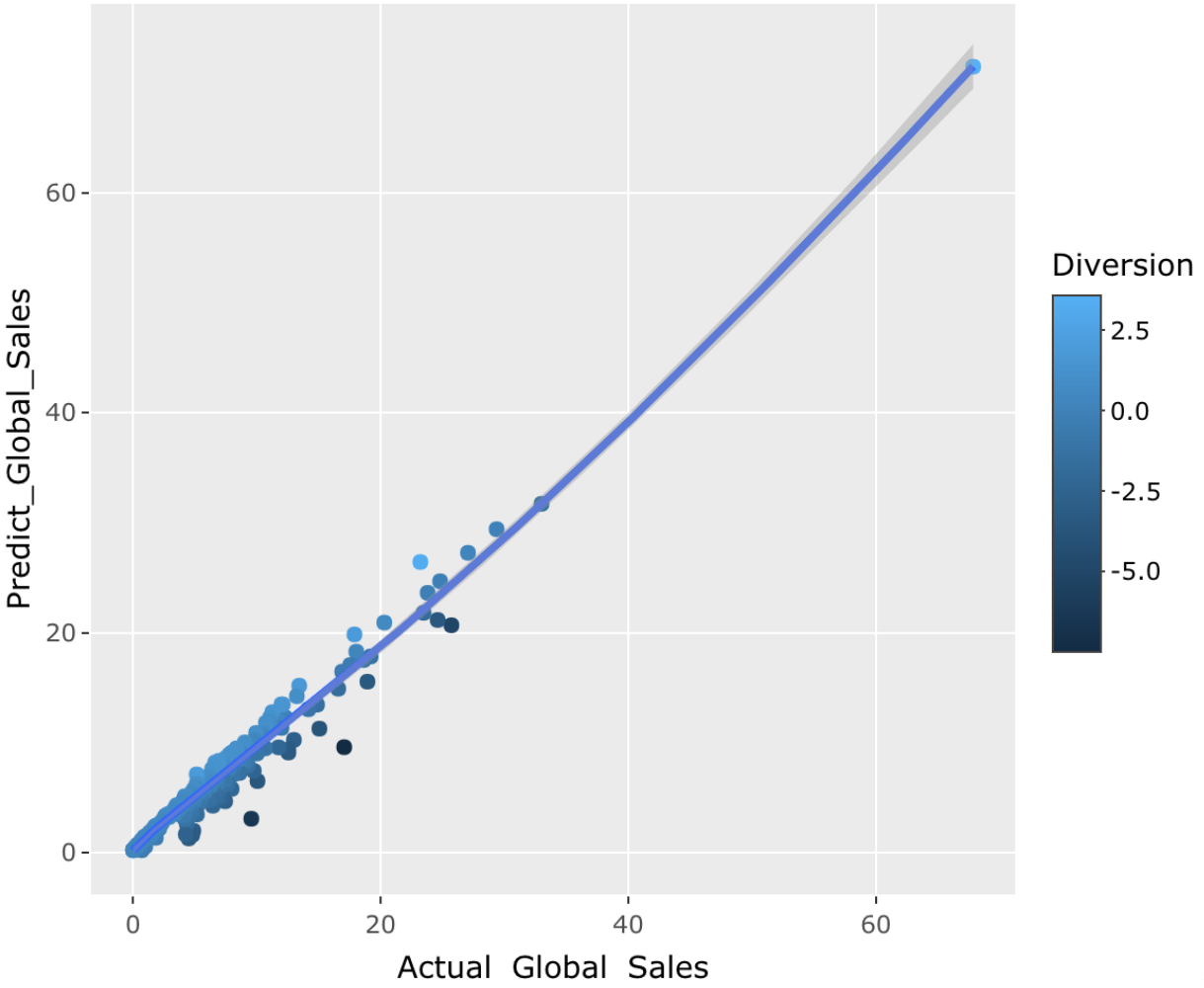
Built on the correlation finding, I built a lrm predicting global sales. After many trials with simple lrm, I decided to move forward with multiple lrm which takes NA and EU sales as independent variables and Global sales as dependent variable. R-squared of the model is high, 96.85%

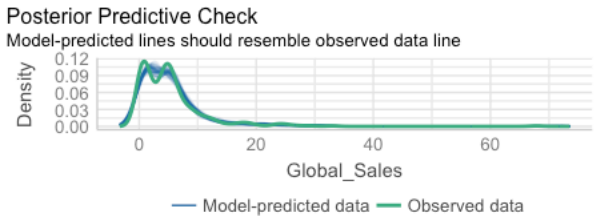
```
Call:
lm(formula = Global_Sales ~ EU_Sales + NA_Sales, data = turtle_sales4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6186 -0.4234 -0.2692  0.0796  7.4639

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22175    0.07760   2.858  0.00453 **
EU_Sales     1.34197    0.04134  32.466  < 2e-16 ***
NA_Sales     1.15543    0.02456  47.047  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.112 on 349 degrees of freedom
Multiple R-squared:  0.9687,    Adjusted R-squared:  0.9685
F-statistic:  5398 on 2 and 349 DF,  p-value: < 2.2e-16
```

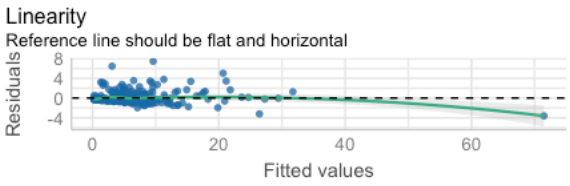
To test the model, first I fed 5 different sales data for EU and NA Sales data to model and compare predicted vs actual values.

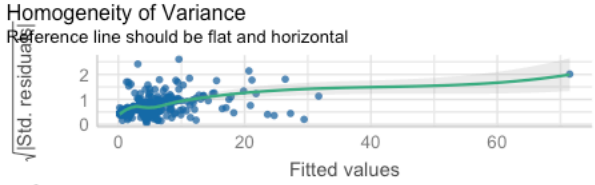
```
  NA_Sales EU_Sales Predict_Global_Sales Actual_Global_Sales Diversion
1    34.02    23.80            71.468572               67.85      3.62
2     3.93     1.56             6.856083                6.04      0.82
3     2.73     0.65             4.248367                4.32     -0.07
4     2.26     0.97             4.134744                3.53      0.60
5    22.08     0.52            26.431567               23.21      3.22
```

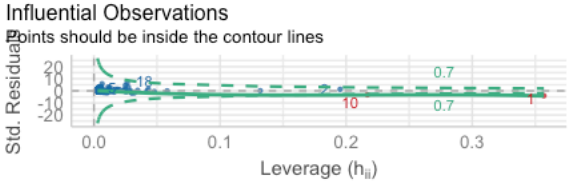
Later, I rerun my model with all data provided (EU & NA sales) and check the performance with easy stat library and scatter plot I created with actual and predicted values. Model prediction works well although there are some prediction errors especially on outlier zones.

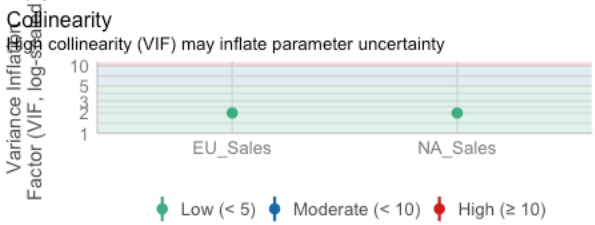

Predict_Global_Sales
Actual Global Sales
Diversion
2.5
0.0
-2.5
-5.0
0
20
40
60

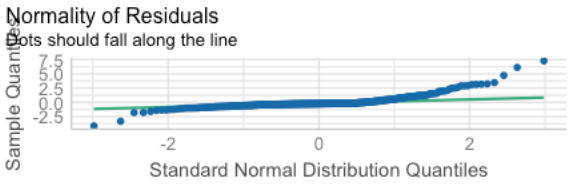

Posterior Predictive Check
Model-predicted lines should resemble observed data line
Density
Global_Sales
Model-predicted data
Observed data


Linearity
Reference line should be flat and horizontal
Residuals
Fitted values


Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
Fitted values


Influential Observations
Points should be inside the contour lines
Std. Residuals
Leverage (h_ii)


Collinearity
High collinearity (VIF) may inflate parameter uncertainty
Variance Inflation Factor (VIF, log-scaled)
EU_Sales
NA_Sales
Low (< 5)
Moderate (< 10)
High (≥ 10)


Normality of Residuals
Dots should fall along the line
Sample Quantiles
Standard Normal Distribution Quantiles

CONCLUSION

Turtle sales company has sales over NA and EU where both regions success in sales contribute to Global sales. There are 5 customer segments based on spending and remuneration. Largest customer group for the company is average earner & average spenders.

General sentiment over products is positive. People mention products/company with good words. Polarity score is positive. Positive word of mouth can be a tool for the company. To increase focus more on this, company can decide to evaluate NPS and include it into its KPIs.

As a further study, I believe there is a room for me to deep dive on product level more. I already identified most purchased products for each customer segment. I can check their sentiments specifically and see what the most liked features are to provide recommendations to business.

As another improvement, negative sentiment received products can be identified and then check whether they are coming from a certain product(s). If that's the case, business can decide to improve the product based on comments or discontinue.

Lastly, the time component on my analysis is missing. It would be better to analyse further to check sales amount and product preference changes over quarters/years to make better predictions.